Home Page

https://www.kaggle.com/competitions/sentiment-analysis-on-movie-reviews/leaderboard

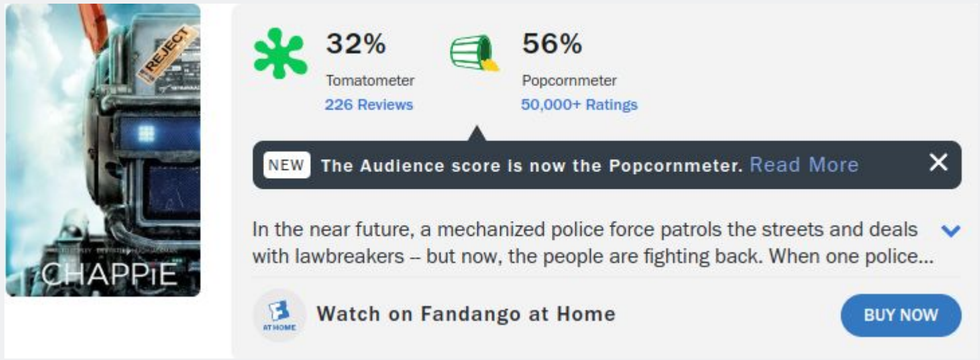
Sentiment Analysis on Movie Reviews

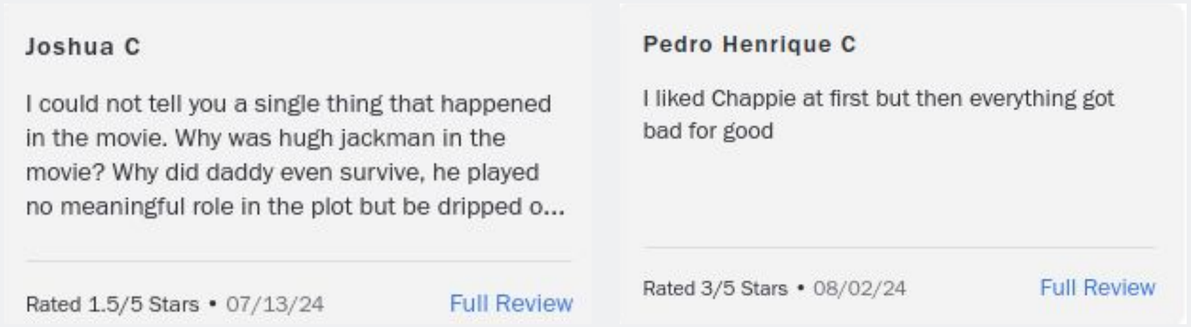
Predictive analytics with machine learning

by Tiago Pedro

Competition Description

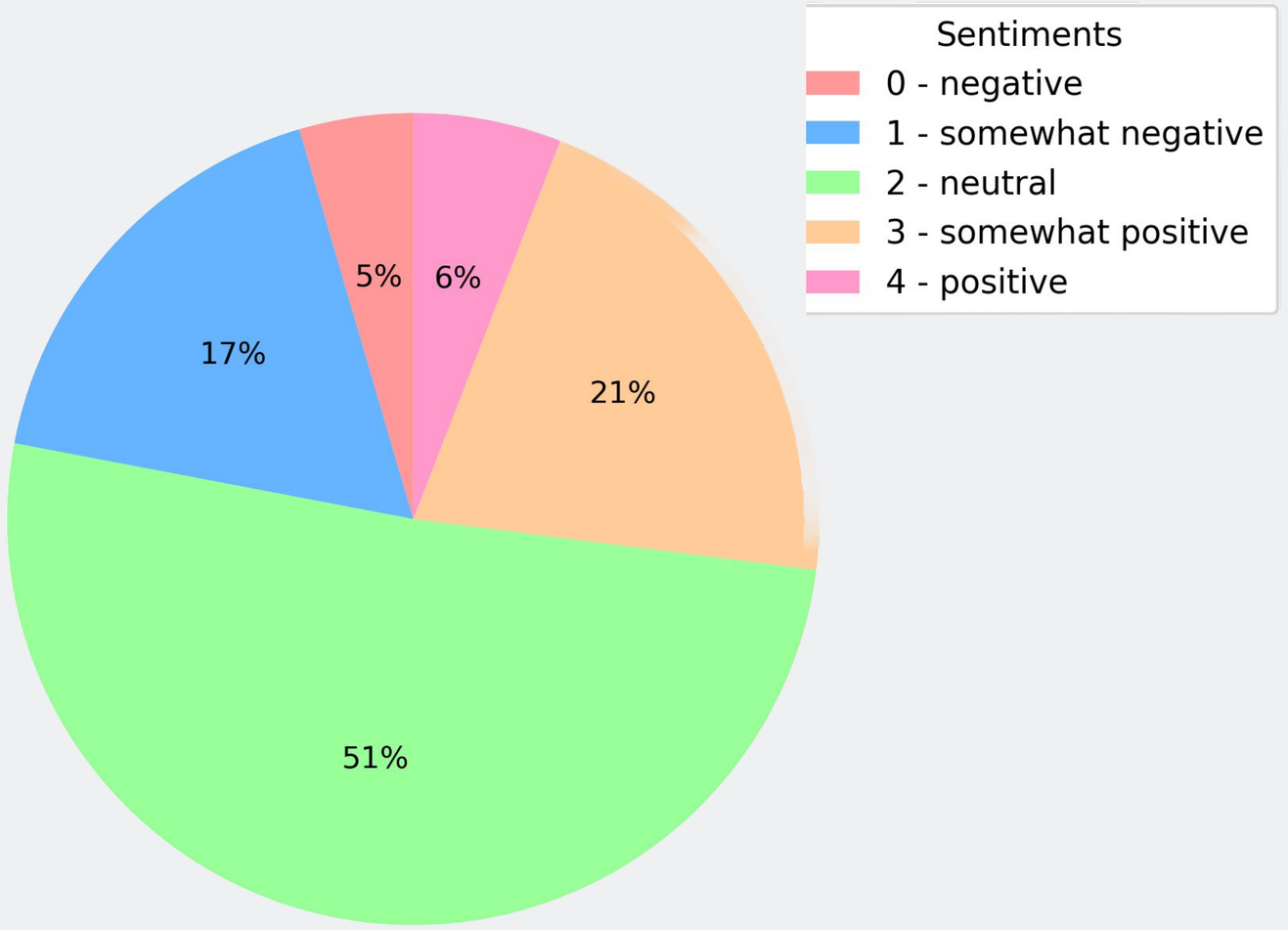
For this competition we are delivered a dataset from Rotten Tomatoes, with comments and the sentiment associated with the comment, we have a training data set and a test data set for the submission.





EDA

The train dataset is composed of 4 columns, ‘PhraseId’, ‘SentenceId’, ‘Phrase’ and ‘Sentiment’, it has 156060 rows, with no duplicated or null values, and the sentiment is a integer that ranges from 0 to 4.



EDA

‘PhraseId’

An integer with the id of the phrase, a phrase then is split in multiple sentences

‘SentenceId’

An integer with the number of the sentence

‘Phrase’

The string that we will use to train the model, containing the comment or parts of it

Data Preprocessing

Null values

The data set had no Null values

SMOTE

Because of the high count of Neutral sentiments, we balanced the dataframe with SMOTE

Spacy

To preprocess the text, including tokenization, lemmatization, and removal of stopwords.

TF-IDF

Term Frequency-Inverse Document Frequency, to balance the importance of the words and convert to a array of numbers

Model Evaluation & Metrics

Model	Accuracy	MAE	RMSE	R2 score
Logistic Regression	0.57	0.49	0.80	0.66
Random Forest	N/A	0.61	0.88	0.06
SVM	0.65	0.40	0.72	0.66
KNN	0.52	0.59	0.90	0.71
XGBClassifier	0.55	0.53	0.53	0.63

Hyperparameter Tuning

Logistic Regression

C (Inverse of Regularization Strength)
penalty (Regularization Type)
solver (Algorithm to Optimize the Model
max_iter (Maximum Iterations)

SVM

C (Regularization Parameter)
kernel (Kernel Function)
gamma (Kernel Coefficient for ‘rbf’, ‘poly’, and ‘sigmoid’)

Final Model

For time purposes i used the Logistic Regression, with 'C': [10], 'penalty': ['l2'], 'solver': ['liblinear'], as the SVM Hyperparameter Tuning is a long process when started it would not have time or computing power to work with it.

```
param_grid_lr = {
    'C': [10],
    'penalty': ['l2'],
    'solver': ['liblinear']
}
```

Playground Prediction Competition - Sentiment Analysis on Movie Reviews

Kaggle Submission

YOUR RECENT SUBMISSION

submission.csv
Submitted by Tiago "ShaQ" Pedro · Submitted a minute ago

Score: 0.56757

↓ Jump to your leaderboard position

569	Vagabond		0.56902	4	10y
570	Abhinav Unnam		0.56893	10	10y
571	Md. Mosharaf Hossain		0.56750	2	10y
572	Daniel B.		0.56658	3	10y

Key Insights & Challenges

Machine Learning Models

Choosing the right ones for the project and the right way to use the parameters

Tokenization

The way we can compile a text to different type of values to input on the ML model

Computing power struggle

Some ML models take longer to train or are even impossible to use with the current computing power

Hyperparameter Tuning

Time consuming to try various ML, the need to test just the right from the ‘standart’ tests

Possible Improvements

- Use of Transformers-Based models like BERT/DistilBERT
- Use of Deep Learning Models like RNNs or CNNs
- More computing power to be able to use the models with the full data set
- Test with different Tokenizers like KERAS
- Test different word embeddings like GloVe or Word2Vec

Conclusion

Machine learning is not only about the model we chose, but the way we manage the data before training the models, in this case using text to predict a data, there are various ways to treat them with different libraries for Tokenization, the lower score on the competition show the need to improve the understanding of these these libraries to optimize the data for better results.

Thank you for your time

Please ask any questions you may have

Project made by Tiago Pedro